FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
SVEUČILIŠTE U ZAGREBU

SEMINARSKI RAD

Kompleksne Mreže

Primjena kompleksnih mreža u proučavanju proteinskih interakcija

Petar Vidaković

Mentor: *Mr. sc. Mile Šikić*

Zagreb, 2007.

Primjena kompleksnih mreža u proučavanju proteinskih interakcija

Sažetak:

Ovaj seminar istražuje kako se korištenjem kompleksnih mreža mogu analizirati proteinske interakcije. Analiza je provedena nad sedam vrsta (S. cerevisiae, H. pyroli, E. coli, C. elegans, H. sapiens, M. musculus i D.melanogaster). Nekoliko ključnih topoloških parametra, (kao što su povezanost čvorišta i prosječnog srednjeg najkraćeg puta) je korišteno kako bi se okarakterizirala mreža proteinskih interakcija (PIN – engl. protein interaction network). Logaritam distribucijske povezanosti u ovisnosti logaritma broja veza indicira da mreže proteinskih interakcija slijede ponašanje zakona potencije (*P(k) ~ $k^{-\gamma}$*)**.** *Utvrđeno je da* γ[1] *leži između 1.5 i 2.4, za svih sedam vrsta. Korelacijska analiza daje dobre dokaze koji podupiru činjenicu da svih sedam PINa formiraju mrežu bez skale (engl. scale-free)*[7]. *Prosječni promjeri mreža i njihove slučajno odabrane verzije su pokazale veliko odstupanje. Također je pokazano da su te interakcije prilično snažne kada izložene nasumičnom poremećaju. Prosjek povezanosti korelacija čvorišta podržava prethodne rezultate da su čvorišta niske povezanosti međusobno povezana, dok čvorišta visoke povezanosti nisu izravno povezana. Ovi rezultati su dali nekoliko dokaza koji sugeriraju da bi takve korelacije mogle biti opće svojstvo PINa uzduž različitih vrsta.*

Ključne riječi: kompleksne mreže, proteinske interakcije, teorija grafova

[1] Eksponent distribucije

Uvod

Proteini su makromolekule koje čine 18-20% našeg tijela. U našim stanicama postoje tisuće različitih proteina aktivnih u bilo koje doba. Mnogi djeluju kao enzimi, katalizirajući kemijske reakcije. Izravno komunicirajući jedan s drugim, neprestano utječu na svoje funkcije.
Nalaze se u krvi, mišićima, koži, kostima, i stalno se izmjenjuju jer se razgrađuju i ponovo sintetiziraju. Brzine kojima se ovi procesi događaju ovisi o vrstama proteina, kako oni komuniciraju jedni s drugima i kako oni komuniciraju s genima. Proteini koji vežu DNK ili RNK često imaju izravan učinak na proizvodnju ili degradaciju drugih proteina.

Proteinske se interakcije u zadnjih par godina promatraju kao složene mreže. Takve složene mreže je moguće naći posvuda u stvarnom svijetu. Internet je mreža kompjutora, mozak je mreža živčanih stanica, a ljudi međusobno komuniciraju kroz mreže odnosa kao što su obitelj, prijatelji ili kolege. Kompleksne mreže donose znanstvenu revoluciju novog milenija[3]. Istraživanje mreža potaklo je veliko zanimanje među biolozima, fizičarima, ekonomistima, informatičarima i sociolozima. Ovaj seminar istražuje kako primjena kompleksnih mreža može koristiti u proučavanju proteinskih interakcija.

U poglavlju „Vrste bioloških mreža“ opisuju se tri velike mreže koje nadgledaju stanične procese. Zatim, u poglavlju „Metode“ opisuje se topologija kompleksnih mreža te se analiziraju rezultati interakcija. U zaključku se razmatra budućnost kompleksnih mreža u biologiji.

Vrste bioloških mreža

Mreže interakcija su temelji svih bioloških procesa; na primjer, stanica može biti opisana kao složena mreža kemikalija spojenih kemijskim reakcijama. Stanični procesi su nadzirani različitim vrstama biokemijskih mreža [5]:

- (i) metabolička mreža
- (ii) mreža interakcija između proteina
- (iii) mreža genskih regulacija.

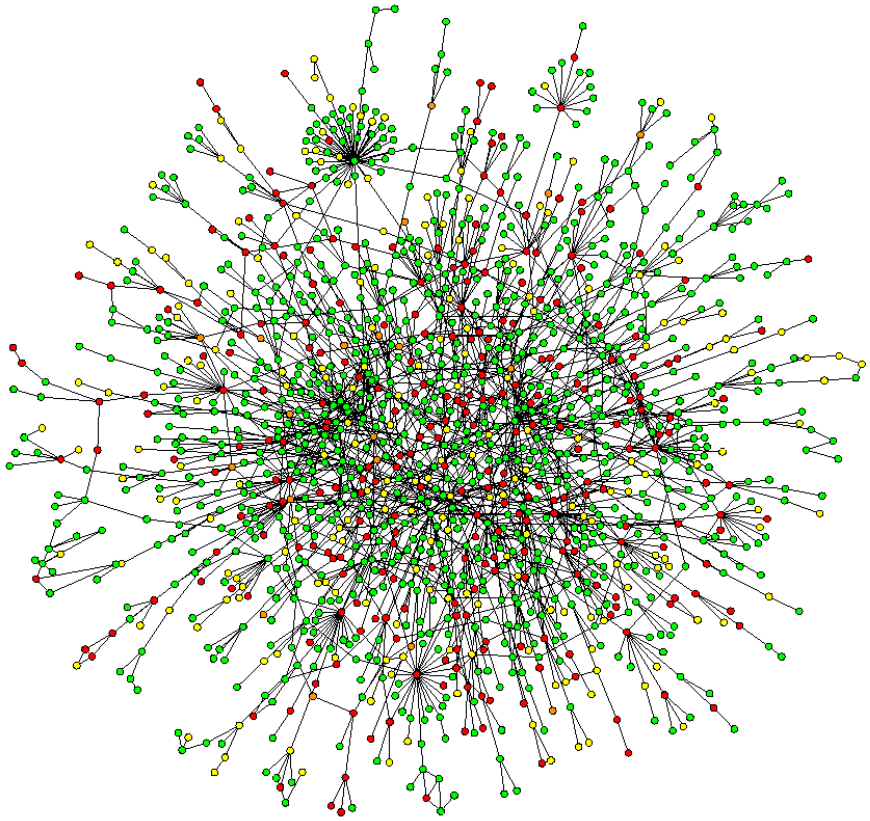
U zadnjih par godina, možemo pratiti napretke u analiziranju bioloških mreža korištenjem statističkih modela kompleksnih mreža. Taj mrežni pristup postao je moćan alat za istraživanje različitih bioloških sustava, kao na primjer proteinskih interakcija u kvascu[6], prehrambenih i metabolička mreža. Istraživanja pokazuju da tamo leže globalne strukture tih bioloških mreža. Niže je prikazan trenutni status ovih istraživanja.

(I) **<u>Metabolička mreža</u>**

Metabolizam obuhvaća mrežu interakcija koja daje energiju i komponente za stanice i organizme. U mnogim kemijskim reakcija u živim stanicama, enzimi djeluju kao katalizatori u konverziji određenih spojeva (supstrati) u druge spojeve (proizvodi). Nedavno, veliki opsega metaboličkih mreža sačinjen od 43 organizama je istraživan i nađeno je da svi imaju svojstvo malih svjetova (engl. *small-world network*) i mreža bez skale, tj. $P(k) \sim k^{-\gamma}$, gdje je k broj veza, te da je promjer metaboličke staze isti za 43 organizma.

(II) **<u>Mreža proteinsko-proteinskih interakcija</u>**

Proteini vrše različite, dobro definirane funkcije, ali malo se zna oko toga kako je njihova interakcija definirana na staničnoj razini. Nedavno je otkriveno da u kvascu (ukupno 329 proteina), proteinsko-proteinske interakcije nisu slučajne, već dobro organizirane. Otkriveno je da većina susjeda visoko povezanih proteina imaju par susjeda, odnosno da su male šanse postojanja interakcija među njima.



Slika I. Mreža interakcija proteina u kvascu

(III) **<u>Regulacija transkripcije gena</u>**

Genetička regulatorna mreža se sastoji od skupa gena i njihove obostrane regulatorne interakcije. Interakcije proizlaze iz činjenice da genetski kod za proteine može kontrolirati izražavanje drugih gena tako da npr. aktiviraju ili inhibiraju transkripciju DNK.

Metode: Topologija složene mreže

Biološke mreže gore navedene imaju složenu topologiju. Složena mreža može biti okarakterizirana određenim topološkim mjerenjima. Erdös i Rényi[2] su prvi predložili model složene mreže poznate kao slučajni graf. Uporabom teorije grafova, svaki protein je predstavljan kao čvorište a interakcija kao veza.

Analiziranjem DIP[3] baze podataka, možemo izgraditi matricu interakcija koja predstavlja PIN. U matrici, određene veličine su pridodijeljene proteinima ovisno o tome da li su im interakcije direktne ili ne.

Tablica 1 pokazuje broj proteina i njihove interakcije za sedam različitih vrsta; *S. cerevisiae H. pylori, C. elegans, E. coli, H. sapiens M. musculus* and *D. Melanogaster.*

Tablica 1. Statistika iz DIP baze za *S. cerevisiae, H. pylori, E. coli, C. elegans, H. sapiens, M. musculus and D. melanogaster*.

Vrsta	Proteini	Interakcije
S. cerevisiae (CORE)	2631	6558
S. cerevisiae	4773	15444
H. pylori	710	1420
E. coli	429	516
C. elegans	2638	4030
H. sapiens	946	1129
M. musculus	324	283
D. melanogaster	7066	21017

Distribucija povezanosti i korelacijska analiza

Prvo topološko svojstvo složene mreže je njen stupanj povezanosti. Iz matrice interakcija možemo dobiti histogram od *k* interakcija za svaki protein. Podjela svake točke histograma s brojem ukupnog broja proteina daje *P(k).* U slučajnoj mreži, veze su nasumično spojene i većina čvorišta ima stupnjeve blizu <*k*>. Stupanj distribucije *P(k)* nasumične mreže s *N* čvorišta, može približno biti određen Poissonovom distribucijom[1], tj. $P(k) \approx e^{-\langle k \rangle} \frac{\langle k \rangle^K}{k!}$ za velik *N*. U mnogim mrežama u stvarnome životu, stupanjska distribucija nema dobro definiran vrhunac ali ima polinomsku distribuciju (engl. power-law), $P(k) \sim k^{-\gamma}$, gdje je γ konstanta. Takve mreže su poznate kao mreže bez skale. Polinomska distribucija

[2] 1959 su prvi puta definirali pojam slučajnih grafova u njihovom djelu "On Random Graphs"

[3] Database of Interacting Protein- baza koja sadrži eksperimentalno određene proteinske interakcije

podrazumijeva da su mreže vrlo nehomogene. U mrežama bez skale, postoji mnogo čvorišta s par veza i nekoliko čvorištima s mnogo veza. Visoko povezani proteini bi mogli igrati ključnu ulogu u funkcionalnosti mreže. Kako bi se izmjerio odnos između P(k) i $k^{-\gamma}$, primjenjuje se metoda korelacije i metoda regresije za analizu log P(k) vs Log k zapisa, i izračuna Pearsonov koeficijent r, koeficijent utvrđivanja r^2, i koeficijent regresije γ .

Shannonova Entropija

Kako bi se kvantitativno okarakterizirala čvorišta povezanosti, koristi se Shannonova entropija koja daje preciznu definiciju informacijske slučajnosti. Uzmimo u obzir binaran niz X duljine *n*, s elementom *xi* koji ima dva stanja, 0 ili 1, Shannonova entropija dobije se formulom

$$H(X) = -\sum_{i=1}^{n} pi \log pi$$

gdje je *pi* vjerojatnost uočavanja 0 ili 1 u određenom nizu, a *n* ukupan stupanj slobode.

Duljina puta interakcije, D

Proteini mogu međusobno imati izravne ili indirektne interakcije. Izravne interakcije kao na primjer spojne interakcije, uključuju oblikovanje proteinskih kompleksa. Indirektna interakcija se odnosi na dva proteina koja neizravno komuniciraju preko uzastopne kemijske reakcije. Još jedna kategorija indirektne proteinske interakcije su ekspresija gena, gdje je poruka od jednog proteina prenošena prema drugom proteinu preko procesa proteinske sinteze.
Drugo topološko mjerenje je udaljenost između dvaju čvorišta, koja se dobije brojem veza po najkraćem putu. Broj veza kojim je čvor povezan s drugim čvorištima se razlikuje od čvora do čvora. Promjer mreže, poznat i kao prosječna duljina puta, je prosjek udaljenosti između svih parova čvorišta. Za sve parove proteina, najkraći put interakcija, j (tj. najmanji broj reakcija kojim se može dosegnuti protein 2 od proteina 1) je određen korištenjem Floydovog algoritma. To je algoritam koji pronalazi najkraći put za svaki vrh u grafu. Algoritam predstavlja mrežu koja ima *N* čvorišta kao *N x N* matricu *M*. Svaki unos (i,j) daje udaljenosti dij od čvora i do čvora j. Ako je i izravno povezan sa j, tada je d_{ij} konačan, a inače beskonačan.

Floydov algoritam se temelji na ideji da, ako su dana tri čvorišta i,j i k, brže se dostiže j od i kada se prolazi kroz k ako
$d_{ik} + d_{kj} < d_{ij}$.

Prosječan promjer, *d*, mreže je zadan kao $d = \frac{\sum_j if(j)}{\sum_j f(j)}$

gdje je j najkraći put, i f(j) frekvencija čvorišta koja imaju duljinu j.

Povezanost čvorova

U modelima slučajnog grafa s proizvoljnom stupanjskom distribucijom, stupnjevi čvorišta su nepovezani. Kako bi ispitali odnose povezanosti čvorišta, može se izbroji frekvencija $P(K_1,K_2)$ povezanih međusobno i usporedit sa istim ali izmjeren kod slučajno odabrane verzije iste mreže. Prosječna povezanost čvorišta (K_2) za fiksan K_1 dobije se relacijom,

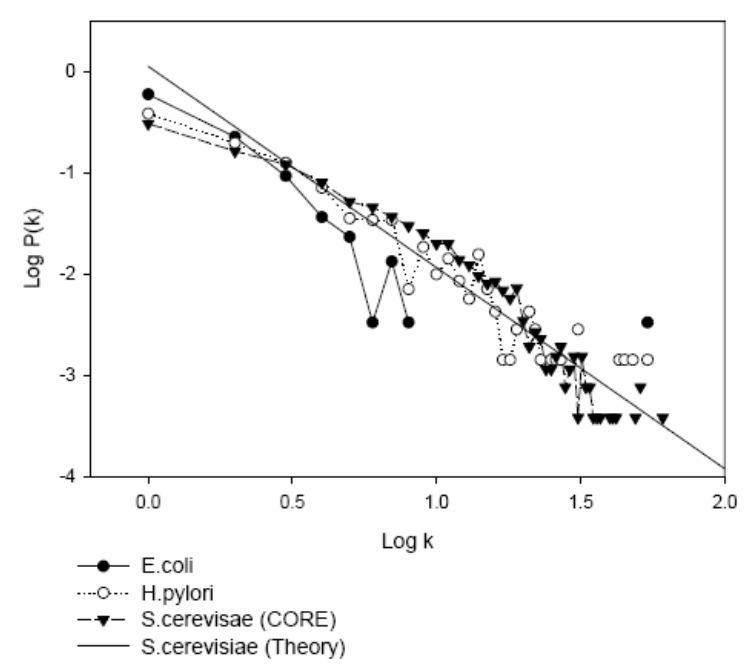
$$\langle K_2 \rangle = \sum K_2 \frac{P(K_1, K_2)}{\langle P_R(K_1, K_2) \rangle}$$

U slučajno odabranoj verziji, povezanost čvorišta svakog proteina je držana isto kao u izvornoj mreži, dok je njihov spojni partner odabran sasvim nasumično.

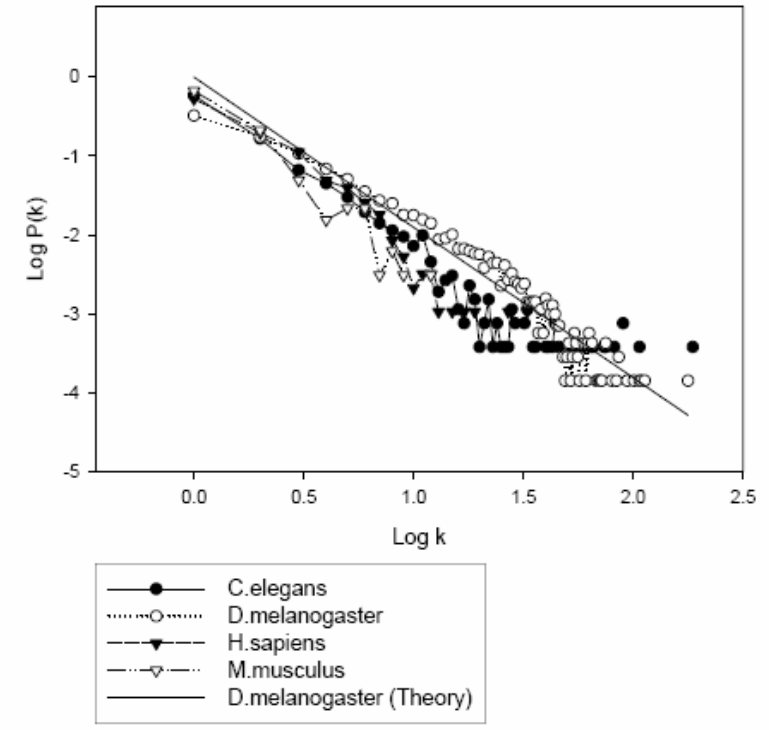
Tablica 2. prosječna povezanost <k>, maksimalna povezanost k_{max}, promjer d, prosječan slučajni promjer $<d_{rand}>$, i Δ rezultat za sedam vrsta.

Vrsta	$<k>$	k_{max}	d	$<d_{rand}>$	$<d_{pert}>$ (Δ)
S. cerevisiae(CORE)	4.87	111	5.01	4.01	5.01 (0.0%)
S. cerevisiae		283	4.19	3.67	4.19 (0.0%)
H. pylori	3.87	54	4.14	3.72	4.14 (0.0%)
E. coli	3.02	54	3.22	5.94	3.40 (5.6%)
C.elegans	1.91	187	4.81	4.43	4.81(0.0%)
H. sapiens	2.30	33	6.05	5.61	6.16 (1.8%)
M. musculus	1.67	12	3.58	6.64	3.74 (4.7%)
D. melanogaster	5.90	80	4.46	5.20	4.46(0.0%)

Prosječan promjer mreža, *d*, je između 3.2 do 6.0 za bilo koja dva proteina. Postoji velika razlika između *d* i $< d_{\text{rand}} >$ za svaku vrstu, zato se može zaključiti da proteinsko-proteinske interakcije vjerojatno nisu slučajne. Također se može zaključiti iz Tablice 2 da su interakcije prilično snažne kada su podložne slučajnim smetnjama, tj. $< d_{\text{pert}} >$ malo odstupa od *d*, odnosno mali iznos delta. To se može interpretirati na način da se proteinsko-proteinska interakcija snažno odupire na vanjske smetnje.



Slika 1. Logaritam normalizirane frekvencijske povezanosti nasuprot logaritma povezanosti za jednostanične organizme, *E.coli, H.pylori,S. carevisae(CORE) i S.cervisiae(CORE,Theory).*



Slika 2. Logaritam normalizirane frekvencijske povezanosti nasuprot logaritma povezanosti za jednostanične organizme, *C. elegans, D. melanogaster, H. sapiens, M. musculus and C. elegans(Theory).*

Očito je iz slika da se broj proteina smanjuje s rastućim brojem veza, tj. da imaju inverznu relaciju.

Logaritam distribucijske povezanosti nasuprot logaritma broja veza pokazuje da PINovi slijede zakon potencije ($P(k) \sim k^{-\gamma}$). On sugerira da PINovi formiraju mreže bez skale.

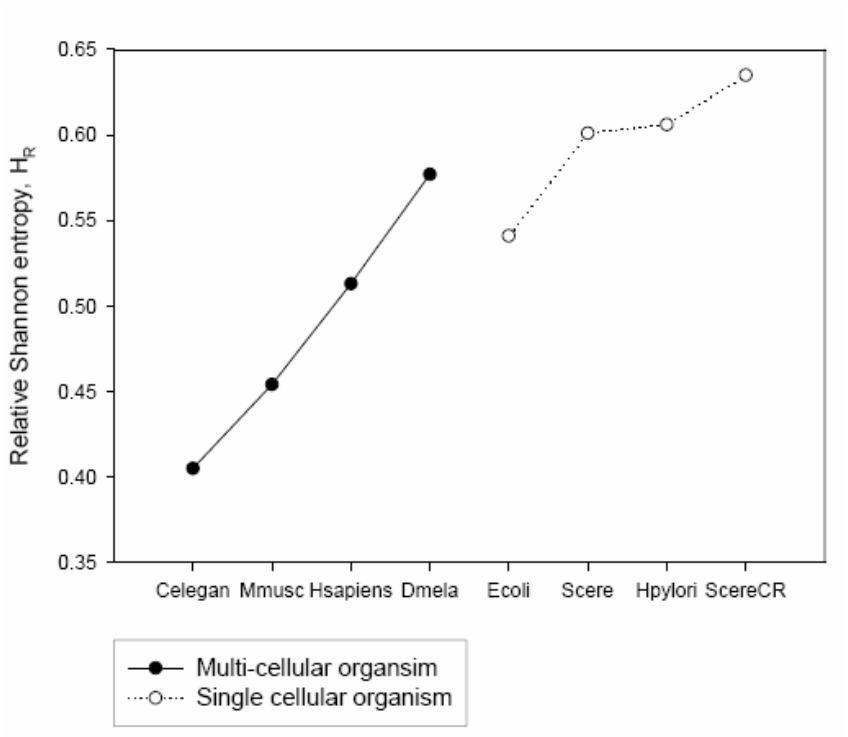
U Tablici 3 je predstavljena regresija i korelacijska analiza za distribucijsku povezanost čvorova. Pomoću regresijske analize, utvrđeno je da je γ između 1.5 i 2.4, za sedam vrsta uzete u obzir.

Korelacijska analiza daje dobre dokaze koji podupiru činjenicu da tih sedam PINova formiraju mrežu bez skale. Što je bliži r^2 prema 1, bolja je korelacija.
Relativno prema drugim vrstama, E.coli ima manji r^2, i zato korelacija nije toliko jaka kao kod drugih vrsta.

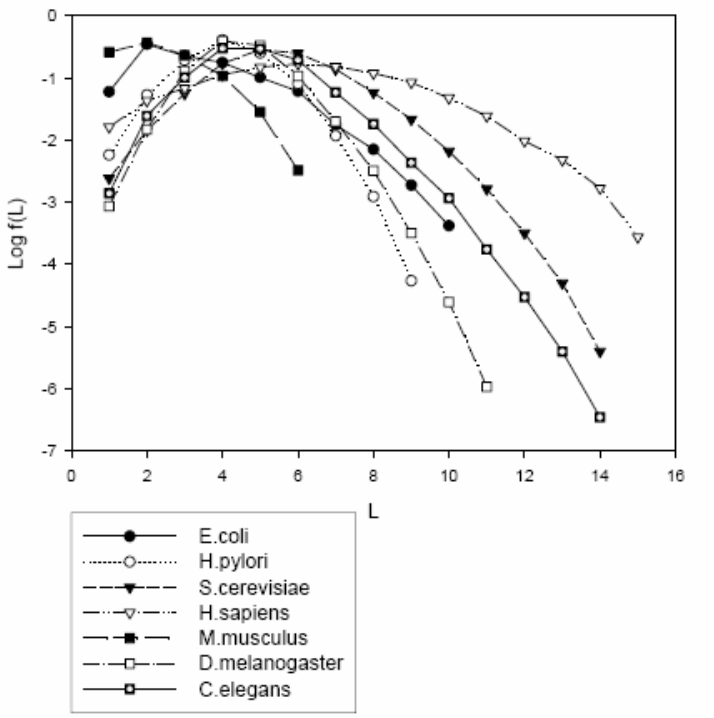
Tablica 3. Eksponent distribucije γ, Paersonov koeficijent r, koeficijent determinacije r^2, ukupan stupanj slobode n, relativna Shannonova entropija HR za bazu DIP.

Vrsta	γ	*r*	r^2	*n*	H_R
S. cerevisiae(CORE)	2.0±0.1	0.95	0.91	44	0.601(0.635)
H. pylori	1.7±0.1	0.95	0.90	30	0.606
E. coli	1.5±0.4	0.84	0.70	9	0.541
C. elegans	1.6±0.1	0.92	0.84	49	0.405
H. sapiens	2.1±0.1	0.96	0.93	19	0.513
M. musculus	2.4±0.2	0.97	0.93	10	0.454
D. melanogaster	1.90±0.02	0.96	0.93	76	0.577

Relativna Shannonova entropija za sedam vrstu je izviještena u Tablici 3 i opisana u Slici 3 također. Entropija je 0.601 i 0.635 za *S. cerevisiae i S. cerevisiae*. Kalkulacija predlaže da višestanični organizmi (izuzevši D. melanogaster) imaju tendenciju manje entropijske vrijednost u usporedi s jednostaničnim organizmima.
Poznato je da je Shannonova entropija mjera nesigurnosti. Što manja entropija (nesigurnost), to je veća struktura uklopljena u podacima.



Slika 3. Relativna Shannonova entropija za jednostanične organizme *(S. cerevisiae, S.cerevisiae (CORE), H. pylori and E. coli)* i višestanične organizme *(C. elegans, H. sapiens, M. musculus and D. melanogaster).*



Slika 4. Logaritam normalizirane frekvencijske distribucije povezanih puteva u ovisnosti logaritma njihovih dužina za *S. cerevisiae(CORE), H. pylori, E. coil, H. sapiens, M. musculus i D. melanogaster.*

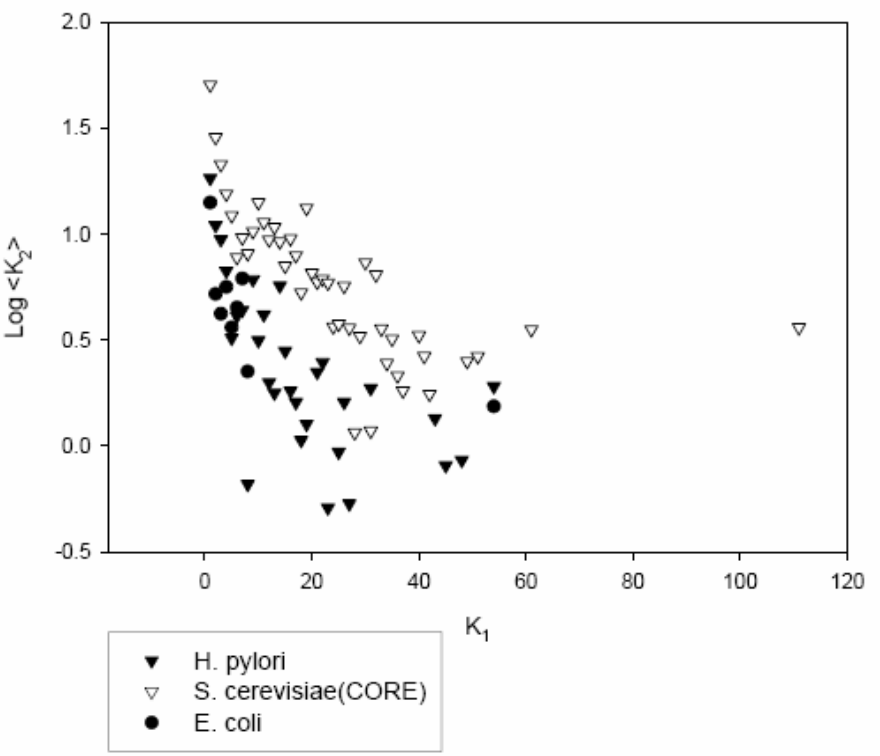
U Tablici 4 je predstavljen najduži put L_{max}, ukupna frekvencija L_{max}-a, f_{max}, i duljina puta s najvišom frekvencijom L' , za sedam vrsta. Nađeno je da *H. sapiens* ima tendenciju većih veličina L_{max} i L' . Što veći L_{max} i L', znači da bilo koja dva proteina mogu imati direktne interakcije preko više uzastopnih kemijskih reakcija.

Tablica 4. najduži put L_{max}, njihove ukupne frekvencije pojavljivanja f_{max}, i duljina puta s najvišom frekvencijom L', za *S.cerevisiae(CORE), H.pylori, E.Coli, C.elegans, H.sapiens, M.musculus i D.melanogaster*.

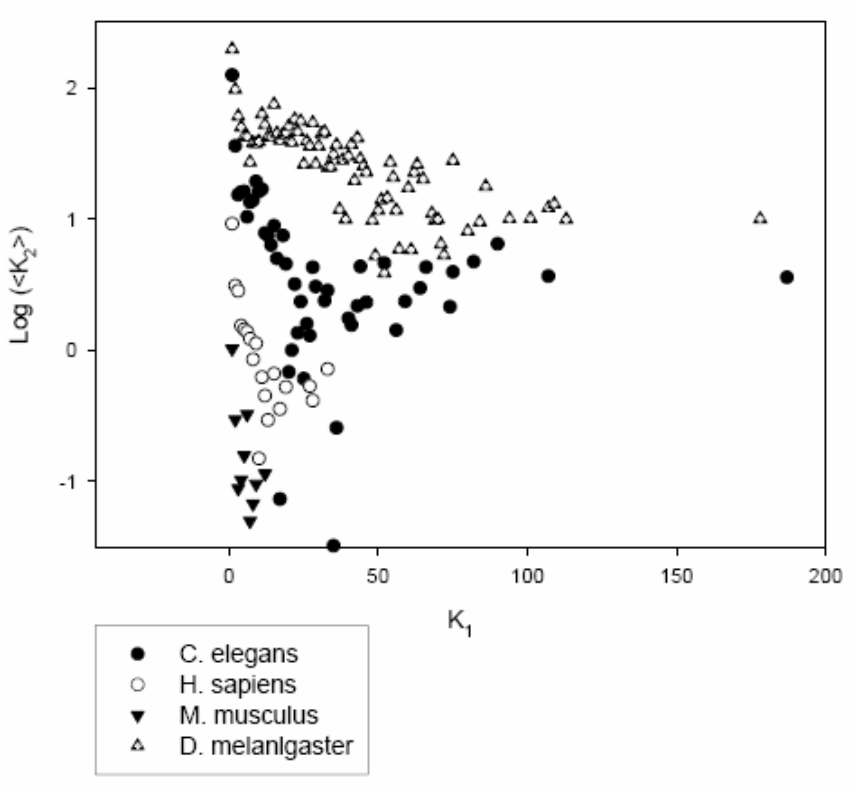
Vrsta	L_{max}	f_{max}	L'(u postocima)
S. cerevisiae(CORE)	13	8	5 (33%)
S. cerevisiae	12	2	4 (45%)
H. pylori	9	26	4 (40%)
E. coli	10	4	2 (35%)
C.elegans	14	2	4(30%)
H. sapiens	16	8	5 (17%)
M. musculus	11	52	4 (39%)
D. melanogaster	14	40	4(19%)

Slike 5 i 6 pokazuju logaritam prosječne stupanjske povezanosti $\log(\langle K_2 \rangle)$ u ovisnosti s povezanosti čvorišta K_1 za jednostanične i višestanične vrste. Ovi rezultati pokazuju da postoji snažna korelacija između čvorišta manjih povezanosti; odnosno gornja lijeva regija svakog nacrta. U nižoj polovica svakog dijagrama postoje par razbacanih točaka koje predlažu da su čvorišta visoke povezanosti veze sa čvorištima manje povezanosti, budući da čvorišta visoke povezanosti nisu izravno povezana.

Ovi rezultati su također dali nekoliko prijedloga da bi korelacijski odnosi mogli biti opće svojstvo PINa uzduž različitih vrsta[7].



Slika 5. Korelacijski prosjek povezanosti čvorova u proteinsko-proteinskim interakcijiama za jednostanične organizme, *H.pyroli, S.cerevisiae(CORE) i E.coli*.



Slika 6. Korelacijski prosjek povezanosti čvorova u protein-proteinskim interakcijama za višestanične organizme, *C.elegans, D.melagonaster, H.sapiens i M.musculus*.

Zaključak

Proučavanjem kompleksnih mreža i njihovih primjena u stvarnome svijetu, znanstvenici su u zadnjih par godina došli do mnogo korisnih informacija kojima se proširuje naše saznanje o životu svih živih organizama.

Možemo se zapitati, što možemo naučiti o biologiji proučavajući mreže? Iako smo još jako daleko od potpunog odgovora na pitanje, možemo ponuditi dva djelomična odgovora. Prvo, takvi pristupi, bazirani na otkrivanju strukture drugih mreža, pomoći će u organizaciji ogromne količine dobivenih podataka i na taj način omogućiti lakši pristup tradicionalnim biolozima. Drugo, redefiniranje postojećih bioloških pitanja iz mrežne perspektive, otvara nam mogućnost obrade velike količine postojećih podataka u cilju odgovara na pitanja koja inače ne bi mogla biti odgovorena. Premda se ove složene mreže predstavljaju teže što više znamo o njima, očekujemo da će naš trud biti obogaćen detaljnom slikom procesa u kojem živa bića razvijaju fenotip od genotipa.
U idućih 5 godina, najvjerojatnije ćemo dobiti čistu sliku proteinskih interakcija. Kada se ti novi podaci upare sa sofisticiranijim mrežama, dobit ćemo predvidljive mreže. Predvidljive mreže će biti korištene u ispitivanju posljedica smetnji na stanice. Također, ovi modeli će vjerojatno biti sposobni, s razumnim stupnjem točnosti, predviđati rezultate ometanja tih interakcija sa postojećim lijekovima. Ako to uspijemo, napraviti ćemo veliki skok u medicini, i dobiti ćemo prijeko potreban alat za razvoj novih lijekova i terapija!

Literatura

[1] S. N. DOROGOVTSEV, J. F. F MENDES: *Evolution of Networks, Oxford*, 2003

[2] S. N. DOROGOVTSEV, J. F. F MENDES: *The Shortest path to complex networks,* 24. srpnja 2004., arXiv:cond-mat/0404593 v4, 27. travnja 2007.

[3] Albert Laszlo Barabasi: *U Mreži*, Naklada Jesenski i TurkČakovec, 2006.

[4] Antoniodel Sol,Hirotomo Fujihashi: *Topology of small-world networks of protein-protein complex structures*,Nature, 2004.

[5] Eric Alm, Adam P Arkin: *Biological networks*, 2003, pp. 193-201

[6] Peer Bork, Lars J Jensten: *Protein interaction networks from yeast to human*, 2003, pp. 292-296

[7] Matteo Pellegrini, *Future drugs*, URL: http://www.future-drugs.com, (2.4.2007)

[8]M.E. Newman: *The structure and function of complex networks*, 2003

Kazalo oznaka i kratica

DIP database	Database of Interacting Protein http://dip.doe-mbi.ucla.edu
PIN	Protein Interaction Network.

Sažetak

U seminaru se definiraju najvažniji pojmovi vezani uz primjenu kompleksnih mreža u proučavanju proteinskih interakcija. Ukratko se definiraju pojmovi najvažnijih mreža. Detaljno se analiziraju rezultati dobiveni primjenom kompleksnih mreža nad sedam vrsta, te uspoređuju interakcije među vrstama.
Pružaju se primjeri i drugih mreža, kao što su metabolička mreža te regulacija transkripcije gena. Proučava se korist primjena kompleksnih mreža, te kako bi nam one mogle doprinijeti u idućih par godina.